Revisiting Disentanglement and Fusion on Modality and Context in Conversational Multimodal Emotion Recognition

Bobo Li
Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University
boboli@whu.edu.cn

Hao Fei
NExT Research Center, School of Computing, National University of Singapore
haofei37@nus.edu.sg

Lizi Liao
School of Computing and Information Systems, Singapore Management University
lzliao@smu.edu.sg

Yu Zhao
Tianjin University
zhaoyucs@tju.edu.cn

Chong Teng
Wuhan University
tengchong@whu.edu.cn

Tat-Seng Chua
National University of Singapore
dcscts@nus.edu.sg

Donghong Ji
Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University
dhji@whu.edu.cn

Fei Li*
Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University
lifei_csnlp@whu.edu.cn

ABSTRACT

It has been a hot research topic to enable machines to understand human emotions in multimodal contexts under dialogue scenarios, which is tasked with multimodal emotion analysis in conversation (MM-ERC). MM-ERC has received consistent attention in recent years, where a diverse range of methods has been proposed for securing better task performance. Most existing works treat MM-ERC as a standard multimodal classification problem and perform multimodal feature disentanglement and fusion for maximizing feature utility. Yet after revisiting the characteristic of MM-ERC, we argue that both the *feature multimodality* and *conversational contextualization* should be properly modeled simultaneously during the feature disentanglement and fusion steps. In this work, we target further pushing the task performance by taking full consideration of the above insights. On the one hand, during feature disentanglement, based on the contrastive learning technique, we devise a Dual-level Disentanglement Mechanism (DDM) to decouple the features into both the modality space and utterance space. On the other hand, during the feature fusion stage, we propose a Contribution-aware Fusion Mechanism (CFM) and a Context Refusion Mechanism (CRM) for multimodal and context integration, respectively. They together schedule the proper integrations of multimodal and context features. Specifically, CFM explicitly manages the multimodal feature contributions dynamically, while CRM flexibly coordinates the introduction of dialogue contexts. On two public MM-ERC datasets, our system achieves new state-of-the-art performance consistently. Further analyses demonstrate that all our proposed mechanisms greatly facilitate the MM-ERC task by making full use of the multimodal and context features adaptively. Note that our proposed methods have the great potential to facilitate a broader range of other conversational multimodal tasks.

CCS CONCEPTS

• **Information systems** → **Multimedia information systems**.

KEYWORDS

Multimodal Learning, Emotion Recognition

ACM Reference Format:
Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. 2023. Revisiting Disentanglement and Fusion on Modality and Context in Conversational Multimodal Emotion Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3581783.3612053

1 INTRODUCTION

The analysis of conversational emotions [47] has received growing attention and has been applied in various downstream tasks, like empathetic response generation [13, 24, 55] and mental disease treatment [48]. Recently, the research on conversational emotion analysis has extended the focus from text to multiple modalities such as video and audio [16, 25, 47]. As illustrated in Figure 1,

*Fei Li is the corresponding author.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612053

Figure 1: An example of multimodal conversation from the MELD dataset [47]. Each utterance comes with three modalities of content: video, audio, and text. The goal of MM-ERC is to recognize the emotion label of each utterance.

multimodal emotion recognition in conversation (named MM-ERC) aims to detect the emotion label for each utterance in a given dialogue by jointly considering auditory, visual, and textual content. The introduction of audio and video compensates for the limitation of solely depending on text features and thus enriches the features used for emotion recognition.

A good number of efforts have been devoted to building effective MM-ERC models and secured promising performance, where the core idea is to effectively disentangle different modalities and then properly fuse them so as to maximize the efficacy of multimodal features for the task [9, 21, 23, 25, 60]. However, MM-ERC intrinsically involves two simultaneous key ingredients: *multiple feature modality* and *conversational contextualization*. While the majority of existing models treat MM-ERC as a typical multimodal classification problem, focusing predominantly on either multimodality or context modeling, the relationship between dialogue context and multimodal feature consistency is often neglected. By revisiting the task of MM-ERC, we note that a sound and effective MM-ERC system should place proper attention to simultaneously modeling the multimodality and contextualization during the feature disentanglement step and fusion step.

Feature Disentanglement. The purpose of feature disentanglement is to extract the critical features from the original feature spaces and weaken the influence of irrelevant features, since multimodal inputs often contain features unrelated to emotion recognition (e.g., background video and noisy audio). While existing models, such as MISA [20] and FDMER [57], propose sophisticated disentanglement mechanisms for single pieces of utterance, disentangling on the conversational contexts has not been considered. On the one hand, different modality features within one utterance should exhibit similarities because, intuitively, multimodal signals under the same utterance can be semantically consistent in representing an identical emotion. On the other hand, features from different utterances with the same modality share similarities in modality-specific characteristics (e.g., timbre, facial expression, and strong wording), which may seem trivial for other modalities but are useful for recognizing emotions in the specific modality space. Feature disentanglement without effectively considering both the modality level and utterance level will inevitably limit further performance improvement of MM-ERC. Unfortunately, to the best of our knowledge, no existing research explores the disentanglement under these two aspects, indicating a potential research gap.

Feature Fusion. The disentangled features from the above step further need to be properly fused, during which reasonable weights are assigned to maximize the utility of features for emotion prediction. Since different clues in varied modalities serve distinct contributions to the final prediction, fusing features across modalities has been extensively considered in existing MM-ERC studies [9, 10, 21, 25], with many sophisticated methods, such as tensor fusion [60], graph convolutional networks [25], gating mechanisms [21]. However, no controllable weights were utilized in previous works, which may risk one modality dominating the multimodal fusion process [45] and potentially limiting the overall performance. Yet we note that the utterance-level fusion should also receive sufficient attention. Intuitively, it is less necessary to further introduce moderate history utterance contexts for prediction when the multimodal signals within the current utterance have indicated a clear emotion tendency in high consistency. Instead, aggressively feeding all the historical contexts would rather deteriorate the inference. For example, in Figure 1, fully considering all previous dialogue contexts might lead to an incorrect emotion determination for the last utterance as "Sadness". This could happen due to the negative neighbor context (i.e., the emotion of the second-last utterance being "Sadness") and the negative atmosphere conveyed throughout the dialogue. Therefore, properly fusing the features from both multimodal ones and dialogue contexts is non-trivial.

In light of the above observations, in this work, we develop a niche targeting solution, i.e., **DF-ERC** (**D**isentanglement & **F**usion for **E**motion **R**ecognition in **C**onversation), to fill the gaps and help achieve higher performance in MM-ERC. As shown in Figure 2, our system comprises four tiers. First, the raw multimodal inputs of dialogues are encoded into various feature extractors to obtain corresponding features. Then, the feature disentanglement layer performs feature disentanglement, where a **Dual-level Disentanglement Mechanism** (DDM) is proposed. DDM employs contrastive learning [14] to push the feature vectors of different modalities or different utterances away, thereby disentangling features at the modality level and utterance level, respectively. Next, the feature fusion layer performs modality-level and context-level integration, in which we propose a **Contribution-aware Fusion Mechanism** (CFM) and a **Context Refusion Mechanism** (CRM) for multimodal and context fusion, respectively. CFM fuses multimodal features based on the true classification probabilities [11] of each modality as their contributions, where such dynamic weighting advances in more controllable feature coordination. In contrast, CRM flexibly schedules the introduction of historical dialogue contexts into the current utterance via a novel emotion-prototype learning strategy. Specifically, CRM calculates the consistency degree of all modality features within an utterance, where a lower

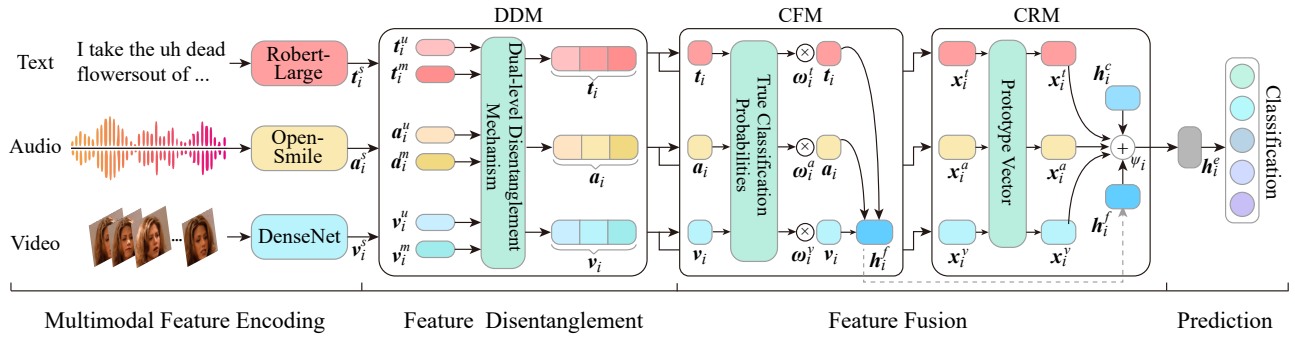


Figure 2: Overall framework of the proposed DF-ERC model. DDM: Dual-level disentanglement Mechanism; CFM: Contribution-aware Fusion Mechanism; CRM: Context Refusion Mechanism.

consistency degree triggers the model to bring in more contexts for reassurance. Finally, the fused overall multimodal and contextual features are used for the emotion label prediction.

To evaluate the efficacy of our proposed approach, we performed extensive experiments on two widely-used benchmarks, namely MELD [47] and IEMOCAP [6]. DF-ERC achieves state-of-the-art performance on overall results and most of the fine-grained emotion categories, demonstrating its effectiveness and stability. Furthermore, we find that DDM was able to effectively disentangle the features of different modalities or utterances (see Figure 10), and the disentangled features played a crucial role in the process of feature fusion (see Figure 6 and Figure 7). Additionally, both CFM and CRM play vital roles, as demonstrated by ablation studies (see Table 2) and in-depth analysis (see Section 4.4). Especially noteworthy is that CRM outperforms the models with no context or full context engagement, demonstrating its superiority (see Figure 8).

Overall, our contributions are four-fold:

- We revisit the MM-ERC task and, for the first time, propose DF-ERC to enhance the task by performing disentanglement and fusion under both the modality and context perspectives.
- Technically, we propose three novel and effective mechanisms to disentangle and fuse both multimodal and contextual features.
- Empirically, our system achieves state-of-the-art performance on two benchmarks.
- Our proposed methods have great potential for facilitating a broader range of conversational multimodal applications.

2 RELATED WORK

Multimodal sentiment analysis [12, 44, 56] aims to extract sentiments or emotions using multiple modality resources, such as text (transcripts), acoustic (audio) and visual modalities. However, discrepancies across different modalities pose a challenge to the model. To address this issue, some studies have focused on modality alignment [52] and minimizing the discrepancies between modalities [39, 59]. Moreover, the style of modality fusion can impact the model performance, leading to the exploration of effective fusion methods, such as hierarchical mutual information [18, 41], reconstruct loss [20], and graph neural network [2, 27, 58]. Additionally, leveraging contextual information to predict dynamic emotions is also a popular approach [1, 7, 8, 15]. However, the use of controllable weights to fuse multimodal features has not been considered in any of the existing approaches, which can limit their performance in practice and is one of the main focuses of our study.

Emotion Recognition in Conversation (ERC) [32, 40, 46] is a subfield of affective computing that aims to recognize emotions for each utterance within a conversation. To develop the model, some studies focus on leveraging dialogue-related features, such as speaker-oriented dialogue modeling[17, 22, 42], context-aware modeling [50, 63], hierarchical feature modeling [31, 34, 35], and emotion transition [4, 51]. With the development of multimodal technology [3, 29, 62], the research scope of ERC has been extended to multimodal scenarios. Many studies have explored multimodal fusion methods for the MM-ERC task, such as multimodal dynamic fusion [21, 36], hierarchical fusion [10], and adaptive modality drop [9]. Although existing adaptive methods have been proposed, they neglect some crucial aspects, such as the contribution of each utterance and the relationship between modality consistency and the involvement of context, which are the main focuses of our paper.

3 FRAMEWORK

Given a dialogue $D = \{u_0, u_1, \cdots, u_n\}$, where u_i represents an utterance, the MM-ERC task aims to recognize the emotion type e_i corresponding to each utterance u_i. In each u_i, there are three kinds of data, namely text, audio, and video, which are used to predict e_i. e_i belongs to a pre-defined set of emotion labels, such as *angry*, *sadness*, *joy*, etc. To approach the task, we introduce a novel framework, termed DF-ERC, illustrated in Figure 2, which performs four tiers of propagation for emotion prediction. Subsequently, we elaborate on the specific techniques employed at each step.

3.1 Multimodal Feature Encoding

Given a dialogue D, we first perform feature extraction for each utterance u_i simultaneously. In this paper, we follow up-to-date previous works [10, 51] and employ RoBerta [37] to obtain contextualized text features. Specifically, all the utterances are concatenated and fed into a pre-trained language model (PLM) following the way in Span-BERT [30]. The dialogue is represented as

$$\begin{aligned} \mathcal{D} &= [[\text{CLS}], w_{11}, w_{12}, \cdots, w_{1l_1}, w_{21}, \cdots, w_{nl_n}, [\text{SEP}]], \\ \boldsymbol{H} &= \text{PLM}(\mathcal{D}), \\ \boldsymbol{t}_i^s &= \text{MeanPooling}(\boldsymbol{H}[start_i, end_i]), \end{aligned} \tag{1}$$

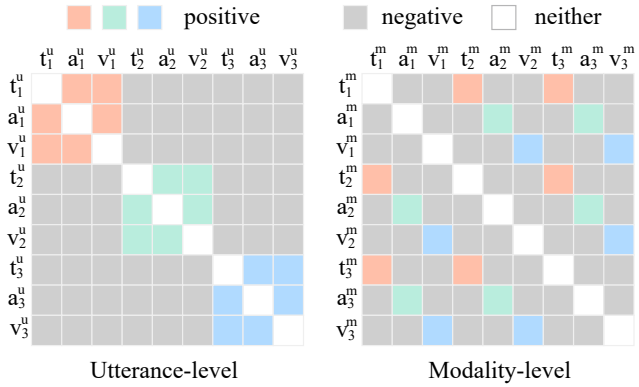


Figure 3: Dual-level disentanglement mechanism, where t_i^*, a_i^*, and v_i^* denote text, audio, and video features, respectively. As seen, utterance-level disentanglement pulls the features of the same utterance close and pushes away the features of different utterances, while modality-level disentanglement pulls the features of the same modality close and pushes away the features of different modalities.

where w_{ij} is the j-th token in u_i and l_i is the length of u_i, $start_i$ and end_i are the indices of the head and tail tokens of u_i in the sequence $\mathcal{D}$, and t_i^s is the text feature for utterance u_i.

For audio and visual content, following the approach described in previous work [21, 25], we adopt OpenSmile [49] and DenseNet [26] pre-trained on the Facial Expression Recognition Plus (FER+) corpus [5] as feature extractors. Finally, we obtain an audio feature a_i^s and a visual feature v_i^s for each utterance u_i.

3.2 Dual-level Disentanglement Mechanism (DDM)

It should be noted that directly utilizing raw multimodal features for emotion analysis is problematic because they are entangled and noisy due to the unconstrained extraction process. Thus, it is necessary to disentangle multimodal features in order to refine them and boost the performance of downstream tasks. In this paper, we propose a dual-level disentanglement mechanism to disentangle raw features in both utterance and modality levels, as illustrated in Figure 3. At the modality level, we apply an MLP layer to the features of different modalities to derive modality-level representations:

$$t_i^m / a_i^m / v_i^m = \mathrm{MLP}^m_{t/a/v}(t_i^s / a_i^s / v_i^s). \tag{2}$$

Next, a list R^m, containing the items t_i^m, a_i^m, v_i^m ($i \in [1, n]$), is created as follows: $R^m = [t_1^m, a_1^m, v_1^m, t_2^m, ..., t_n^m, a_n^m, v_n^m]$. We then apply contrastive learning to these features in order to draw features of the same modality closer to each other and push features of different modalities away, formalized as below:

$$\mathcal{L}_{cl}^m = -\sum_i^n \sum_{h_k \in R_{i+}^m} \log \frac{e^{sim(h_i, h_k)/\tau}}{\sum_{j!=i}^{3n} e^{\mathrm{sim}(h_i, h_j)/\tau}}, \tag{3}$$

where h_i is the i-th item in R^m. Note that $h_k \in R_{i+}^m = \{x_j | j \equiv i (\mathrm{mod}\ 3), 1 < j \leq 3n\}$ has the same modality as h_i, which can be considered as positive instances. Here τ is a temperature parameter, and $sim(h_i, h_k)$ denotes the cosine similarity between two vectors.

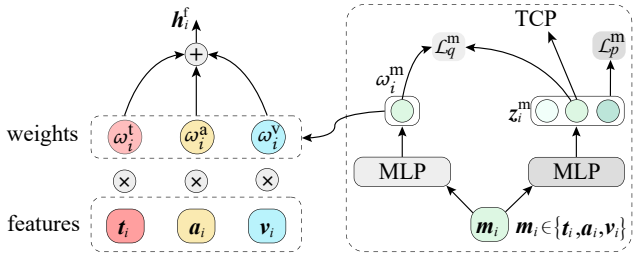


Figure 4: Contribution-aware fusion mechanism. t_i, a_i and v_i are the features from DDM (see Eqs. (6) to (8)). ω_i^t, ω_i^a and ω_i^v are the contributions of different modalities (see Eq. (12)), which are given by three contribution prediction networks MLP_m^p ($m \in \{t, a, v\}$). These networks are trained as students based on true classification probabilities (TCPs) given by three teacher networks MLP_m^q.

At the utterance level, contrastive learning is exploited in a similar manner, clustering the multimodal features of the same utterance and disentangling the features of different utterances, formulated as below:

$$t_i^u / a_i^u / v_i^u = \mathrm{MLP}^u_{t/a/v}(t_i^s / a_i^s / v_i^s), \tag{4}$$

$$\mathcal{L}_{cl}^u = -\sum_i^n \sum_{h_k \in R_{i+}^u} \log \frac{e^{sim(h_i, h_k)/\tau}}{\sum_{j!=i}^{3n} e^{\mathrm{sim}(h_i, h_j)/\tau}}, \tag{5}$$

where $h_i \in R^u = [t_1^u, a_1^u, v_1^u, t_2^u, \cdots, t_n^u, a_n^u, v_n^u]$, and $h_k \in R_{i+}^u = \{x_j | \lfloor i/3 \rfloor = \lfloor j/3 \rfloor, 0 < j \leq 3n\}$ denotes the feature in the same utterance with h_i. $\lfloor x \rfloor$ is the round down symbol. Finally, we use residual connections to concatenate raw features with disentangle features, and two loss functions for contrastive learning are also combined:

$$t_i = [t_i^s; t_i^m; t_i^u], \tag{6}$$

$$a_i = [a_i^s; a_i^m; a_i^u], \tag{7}$$

$$v_i = [v_i^s; v_i^m; v_i^u], \tag{8}$$

$$\mathcal{L}_{cl} = \mathcal{L}_{cl}^m + \mathcal{L}_{cl}^u. \tag{9}$$

3.3 Contribution-aware Fusion Mechanism (CFM)

As different modalities have different importance for the final emotion label prediction, they should be assigned different fusion weights in the modality fusion process. Here we adopt a contribute-aware adaptive fusion module to assign the weight of each modality, which can give a dynamic weight according to their prediction performance, as illustrated in Figure 4. Specifically, we apply a classifier on the representation of each modality and obtain the true classification probability (TCP) [11, 19] as their contribution in the fusion process, which can be obtained via:

$$z_i^{t/a/v} = \mathrm{Softmax}(\mathrm{MLP}^q_{t/a/v}(t_i / a_i / v_i)), \tag{10}$$

$$\mathrm{TCP}_i^m = (z_i^m)_{I_i^*}, m \in \{t, a, v\}, \tag{11}$$

where z_i^m is the prediction probability, and I_i^* is the index of golden emotion label for u_i. Obviously, $\mathrm{TCP}_i^m \in (0, 1)$ denotes how likely the prediction result is right. A larger TCP value indicates the feature representation $t_i / a_i / v_i$ can yield a correct prediction result

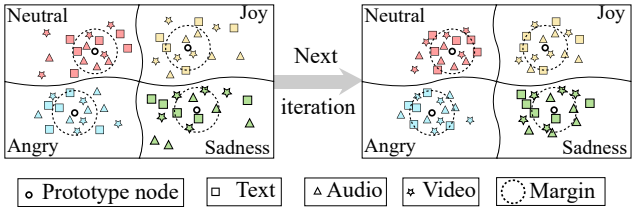


Figure 5: An illustration of the process of prototype-based alignment module.

and verse visa. Therefore, we plan to adopt TCP to represent the fusion weight for each modality. However, a significant challenge arises during the evaluation phase as the true emotion label is unknown, making it impossible to directly utilize TCP as the weight. To keep the consistency of training and test processes, we adopt a predicted value, which is trained to be close to TCP, as the weight of each modality:

$$\omega_i^{t/a/v} = \text{Sigmoid}(\text{MLP}^p_{t/a/v}(\boldsymbol{t}_i/\boldsymbol{a}_i/\boldsymbol{v}_i)). \tag{12}$$

To achieve the goal that we mentioned before, the following loss functions are used:

$$\mathcal{L}_p^m = -\sum_{i=1}^{n} \log(z_i^m(I_i^*)), \tag{13}$$

$$\mathcal{L}_q^m = \sum_{i=1}^{n} \text{MSE}\ (\text{TCP}_\text{i}^\text{m}, \omega_\text{i}^\text{m}), \tag{14}$$

$$\mathcal{L}_{con} = \sum_{m\in\{t,a,v\}} (\mathcal{L}_p^m + \mathcal{L}_q^m), \tag{15}$$

where, $\mathcal{L}_p^m$ represents the prediction label loss and $\mathcal{L}_q^m$ is the prediction TCP loss. Finally, the fused multimodal features can be formulated as:

$$\boldsymbol{h}_i^f = \omega_i^t \boldsymbol{t}_i + \omega_i^a \boldsymbol{a}_i + \omega_i^v \boldsymbol{v}_i. \tag{16}$$

3.4 Context Refusion Mechanism (CRM)

Except for fusing multimodal features, contextual feature fusion is also important, especially when multimodal features contradict each other regarding emotion prediction. Therefore, we compute the agreements among multimodal features as the weights to determine how many contextual features should be incorporated. However, since multimodal features are not aligned according to emotions, it may be inaccurate to directly compute the similarity based on multimodal features. To solve this problem, we propose a prototype-based alignment module (as shown in Figure 5) to learn the emotion-specific representations of multimodal features. Specifically, in each training epoch, we maintain a prototype vector for each kind of emotion:

$$\boldsymbol{x}_i^{t/a/v} = \text{MLP}^r_{t/a/v}(\boldsymbol{t}_i/\boldsymbol{a}_i/\boldsymbol{v}_i), \tag{17}$$

$$R_r^k = \frac{1}{|N_r^k|}(R_{r-1}^k \cdot N_{r-1}^k + \sum_{m\in\{t,a,v\}} \sum_{\text{I}_i^*=k} \boldsymbol{x}_i^m), \tag{18}$$

where R_t^k represents the prototype of the k-th emotion in the t-th epoch, $N_r^k = N_{r-1}^k + 3 \cdot \sum_i^n 1_{I_i^*==k}$ is the size of the k-th emotion in the r-th round. The prototype vector is updated in each iteration based on the previous values and the multimodal features in the current epoch. To ensure that each feature is close to its prototype vector, we use a margin-based loss function based on the mean squared error (MSE):

$$\mathcal{L}_{sim} = \frac{1}{3n}\sum_{i=1}^{n}\sum_{m,k}\sum_{I_i^*==k} max(\text{MSE}(R_r^k, \boldsymbol{x}_i^m) - \beta, 0), \tag{19}$$

where $\mathcal{L}_{sim}$ denotes the loss function and β represents the margin. If the MSE between a feature vector and its corresponding prototype vector is less than the margin, the model will not be updated. In other words, the feature vector is expected to be close to the prototype vector but not necessarily identical to it, in order to avoid all feature vectors becoming indistinguishable. Once the multimodal feature vectors are aligned, the comprehensive similarity between different modalities in the utterance u_i can be computed as:

$$\psi_i = \frac{1}{3}(\cos(\boldsymbol{x}_i^t, \boldsymbol{x}_i^a) + \cos(\boldsymbol{x}_i^t, \boldsymbol{x}_i^v) + \cos(\boldsymbol{x}_i^a, \boldsymbol{x}_i^v)). \tag{20}$$

In the context fusion stage, we first utilize a bidirectional long short-term memory (Bi-LSTM) to generate contextual representations of the utterances as follows:

$$\boldsymbol{h}_1^c, \boldsymbol{h}_2^c, \cdots, \boldsymbol{h}_n^c = \text{BiLSTM}([\boldsymbol{h}_1^f, \boldsymbol{h}_2^f, \cdots, \boldsymbol{h}_n^f]). \tag{21}$$

We then concatenate fused features with context-aware features as follows[1]:

$$\boldsymbol{h}_i^e = [\boldsymbol{h}_i^f, \boldsymbol{h}_i^c \cdot (1 - \psi_i)], \tag{22}$$

where $\boldsymbol{h}_i^e$ represents the final representation of each utterance that is fused with multimodal and contextual information.

3.5 Prediction and Learning

Then, the fused representation $\boldsymbol{h}_i^e$ is used for emotion recognition, which is performed as follows:

$$\boldsymbol{y}_i = \text{Softmax}(\text{MLP}^c(\boldsymbol{h}_i^e)), \tag{23}$$

where $\boldsymbol{y}_i$ represents the probability of predicted emotion. We use the cross-entropy loss function for training, which is defined as follows:

$$\mathcal{L}_{emo} = -\sum_{i=1}^{n} \log \boldsymbol{y}_{i,\text{I}_i^*}, \tag{24}$$

where I_i^* denotes the golden label index of the utterance u_i. During the learning stage, our training loss functions consist of the following parts:

$$\mathcal{L} = \alpha_1\mathcal{L}_{cl} + \alpha_2\mathcal{L}_{con} + \alpha_3\mathcal{L}_{sim} + \mathcal{L}_{emo}, \tag{25}$$

where $\alpha_{1-3} \in (0, 1]$ are hyper-parameters.

4 EXPERIMENTS75

4.1 Implementation Details

Datasets. We conducted experiments using two publicly available MM-ERC datasets, namely MELD [47] and IEMOCAP [6]. Both of which include text, audio, and video modalities. The data split used in our experiments follows previous work [21, 25], and the detailed corpus statistics are presented in Section A.1 of the Appendix.

[1]We also conducted an experiment by directly multiplying $(1 - \psi_i)$ with each $\boldsymbol{h}_*^f$ in Eq. (21) instead of $\boldsymbol{h}_i^c$ in Eq. (22), and obtained a similar result. Therefore, we adopted the more concise fusion approach as shown in Eq. (22).

Table 1: Comparisons with the baselines. 'W-F1' refers to weighted F1 scores. The results with waveline denote the best baseline results. The results with ⋆ denote significance at $p < 0.01$ compared with the best baseline results. The '-' symbol denotes that the corresponding item is not reported in the original paper. Furthermore, the term 'KG' indicates the model is augmented with a knowledge graph.

			IEMOCAP								MELD						
Input	Model	Embedding	Hap	Sad	Neut	Ang	Exci	Frus	Acc	W-F1	Neut	Surp	Sad	Joy	Ang	Acc	W-F1
Text	DiaGCN [17]	Glove	51.57	80.48	57.69	53.95	72.81	57.33	63.22	62.89	75.97	46.05	19.60	51.20	40.83	58.62	56.36
	HiTrans [34]	Bert-Base	-	-	-	-	-	-	-	64.50	-	-	-	-	-	-	61.94
	RGAT [28]	Bert-Base	-	-	-	-	-	-	-	65.22	-	-	-	-	-	-	60.91
	TUCORE [31]	Robe-Large	-	-	-	-	-	-	-	-	-	-	-	-	-	-	65.36
	DiaCRN [22]	Glove	53.23	83.37	62.96	66.09	75.4	66.07	67.16	67.21	77.01	50.10	26.63	52.77	45.15	61.11	58.67
	EmoFlow [51]	Robe-Large	-	-	-	-	-	-	-	-	-	-	-	-	-	-	66.50
T+KG	TODKAT [64]	Robe-Large	-	-	-	-	-	-	-	61.33	-	-	-	-	-	-	65.47
	CoMpM [33]	Robe-Large	-	-	-	-	-	-	-	69.46	-	-	-	-	-	-	66.52
	COMSMIC [16]	Robe-Large	-	-	-	-	-	-	-	65.28	-	-	-	-	-	-	65.21
	SKSEC [53]	Robe-Large	-	-	-	-	-	-	-	66.47	-	-	-	-	-	-	66.21
MM	TFN [60]	Glove	37.26	65.21	51.03	54.64	58.75	56.98	55.02	55.13	77.43	47.89	18.06	51.28	44.15	60.77	57.74
	MFN [61]	Glove	48.19	73.41	56.28	63.04	64.11	61.82	61.24	61.60	77.27	48.29	23.24	52.63	41.32	60.80	57.80
	DiaRNN [42]	Word2vec	32.2	80.26	57.89	62.82	73.87	59.76	63.52	62.89	76.97	47.69	20.41	50.92	45.52	60.31	57.66
	MMGCN [25]	FastText	45.14	77.16	66.36	68.82	74.71	61.04	66.36	66.26	76.33	48.15	26.74	53.02	46.09	60.42	58.31
	MetaDrop [9]	Robe-Large	-	-	-	-	-	-	69.38	69.59	-	-	-	-	-	66.63	66.30
	DiaTRM [43]	Bert-Base	-	-	-	-	-	-	69.50	69.70	-	-	-	-	-	65.70	63.50
	MM-DFN [21]	FastText	42.22	78.98	66.42	**69.77**	75.56	66.33	68.21	68.18	77.76	50.69	22.93	54.78	47.82	62.49	59.46
	M2FNet [10]	Robe-Large$_e$	**60.00**	82.11	65.88	68.21	72.6	**68.31**	69.86	69.69	80.06	58.66	**47.03**	65.5	55.25	67.85	66.71
	UniMSE [23]	T5-Base	-	-	-	-	-	-	70.56	70.66	-	-	-	-		65.09	65.51
Our	DF-ERC	Robe-Large	56.37	**84.36**	**71.13**	67.46	**79.11**	66.23	**71.84⋆**	**71.75⋆**	**80.17**	**60.27**	43.89	**65.93**	55.50	**68.28⋆**	**67.03⋆**

Settings. Following previous work [9, 10, 51], we adopt RoBERTa-large [37] as our PLM encoder, with a hidden state dimension of 1024. For the MELD dataset, we empirically set hyperparameters α_{1-3} to 0.3, 0.8, and 0.3, respectively. For the IEMOCAP dataset, we set the values to 0.2, 0.9, and 1.0. More details about the hyperparameter settings can be found in Section A.2 of the Appendix. We determine all hyperparameters through development experiments on the validation sets. We report the final results as an average over five random seeds, and we consider the evaluation score significant when the p-value is less than 0.01.

Baselines. We compare the performance of our DF-ERC model with several strong baselines, including text-based models, text + knowledge-enhanced models, and multimodal-based models, which are listed in Table 1. We also present the pre-trained embedding weight of each model, most of which use Robert-Large to encode the text content. We present the original results for each baseline, except TODKAT [64] where we use updated results from the paper's repository.[2]. The training datasets in UniMSE [23] are merged from three corpora, possibly contributing to improved performance.

4.2 Main Results

Table 1 presents the experimental results for two benchmark datasets. When compared to text-based models, DF-ERC significantly improves performance scores. Specifically, for the MELD dataset, DF-ERC improves accuracy (Acc) and weighted F1 (W-F1) scores by 7.17 and 0.53 percentage points (hereafter, 'points') as compared to the best-performing baseline, respectively. Similarly, for the IEMOCAP dataset, DF-ERC improves Acc and W-F1 scores by 4.68 and 4.54 points, respectively. These findings underscore the efficacy of incorporating multimodal information and the efficient integration of multimodal features.

Interestingly, without utilizing any external knowledge, DF-ERC still surpasses models that rely on external knowledge, such as TODKAT, CoMpM, COMSMIC, and SKSEC. The table demonstrates that DF-ERC improves the W-F1 scores by 0.51 and 2.29 points for the MELD and IEMOCAP datasets, respectively. These findings suggest that multimodal information effectively compensates for the absence of external knowledge, thus enhancing emotion recognition performance.

Additionally, DF-ERC achieves the best performance among all multimodal-based models. On the MELD dataset, DF-ERC improves Acc and W-F1 scores by 0.43 and 0.32 points, respectively. On the IEMOCAP dataset, the improvements are 1.28 and 1.09 points. These results indicate that DF-ERC is adept at discerning the differences and weighted contributions of each type of multimodal feature. Consequently, DF-ERC optimally utilizes multimodal features to bolster multimodal emotion recognition in conversation.

Lastly, we present the performance scores for each type of emotion, revealing that DF-ERC achieves the best performance for most emotions, thereby demonstrating the robustness of our model. It also contributes to the superior overall performance of our model.

4.3 Ablation Studies

Ablation studies for DDM, CFM, and CRM. As shown in Table 2, we observe that upon the removal of the DDM, utterance-level

[2]https://github.com/something678/TodKat

Table 2: Ablation studies for DDM, CFM, and CRM, where '+Att' denotes the application of a self-attention mechanism for feature fusion, where 'full' and 'zero' means the weight of contextual features (Eq. (22)), i.e., using full contextual features or none of them. The notions 'Utterance' and 'Modality' correspond to the removal of utterance-level and modality-level contrastive learning, respectively.

Model	MELD		IEMOCAP	
	Acc	W-F1	Acc	W-F1
DF-ERC	68.28	67.03	71.84	71.75
- DDM	66.36(↓ 1.92)	65.59(↓ 1.44)	69.99(↓ 1.85)	69.81(↓ 1.94)
- Utterance	66.91(↓ 1.37)	65.92(↓ 1.11)	71.45(↓ 0.39)	70.55(↓ 1.20)
- Modality	67.78(↓ 0.50)	66.48(↓ 0.55)	70.95(↓ 0.89)	70.94(↓ 0.81)
- CFM	66.51(↓ 1.77)	65.49(↓ 1.54)	69.69(↓ 2.15)	69.56(↓ 2.19)
+ Att	65.63(↓ 2.65)	65.87(↓ 1.16)	71.41(↓ 0.43)	71.40(↓ 0.35)
- CRM(full)	64.98(↓ 3.30)	65.07(↓ 1.96)	69.56(↓ 2.28)	69.42(↓ 2.33)
- CRM(zero)	65.06(↓ 3.22)	65.09(↓ 1.94)	69.75(↓ 2.09)	69.71(↓ 2.04)
+ Att	66.70(↓ 1.58)	65.97(↓ 1.06)	70.06(↓ 1.78)	69.34(↓ 2.41)

Table 3: Ablation studies for different modalities, where T, A, and V denote Text, Audio, and Video, respectively.

Model	MELD		IEMOCAP	
	Acc	W-F1	Acc	W-F1
DF-ERC(T+A+V)	68.28	67.03	71.84	71.75
T	65.17	64.54	65.13	65.46
A	43.83	41.72	41.47	38.62
V	46.05	36.65	32.84	22.70
T+V	65.33	64.54	70.61	69.49
T+V	65.10	64.95	66.17	65.89
A+V	48.70	45.00	55.70	55.07

disentanglement, and modality-level disentanglement, there is a decrease in the model's performance on both datasets, indicating feature disentanglement is crucial for emotion prediction. Additionally, the model's performance drops by around one point without the CFM. While an attention-based mechanism does offer some assistance, it remains inadequate when compared with the CFM. We attribute this to the fact that different modalities make varying contributions, and the use of a contribution-aware approach allows for the adaptive learning of weights for different modality features, resulting in a better fusion of multimodal features.

Last but not least, CRM also provides assistance in emotion recognition. After removing the CRM module, we directly set the context weight to 1 (full weight) or 0 (zero weight), signifying the introduction of all or no context information. We observe that these static weights impair the model's performance, resulting in a drop of more than 2 points on both the MELD and IEMOCAP datasets. This finding suggests that context representations are not inherently useful or useless. We ultimately achieve better performance using the CRM to determine the weights of the context.

Ablation studies for modalities. As demonstrated in Table 3, our initial findings reveal that the text-based model outperforms other

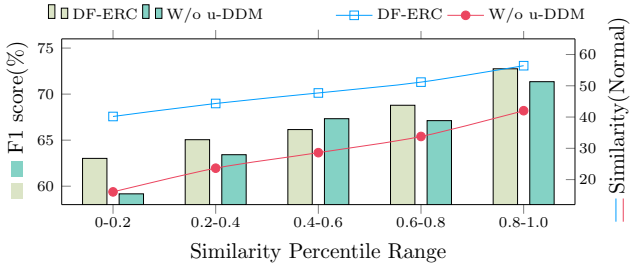


Figure 6: Influence of utterance-level disentanglement (u-DDM) on similarities between modalities and final performance. The x-axis denotes the similarity percentile range, dividing utterances into five groups based on their corresponding similarity values(see ψ_i in Eq. (20)).

modality-based models, providing evidence of the dominance of text as a modality, which is consistent with previous research findings [20]. However, compared to state-of-the-art text-based models in Table 1, our text-based model exhibits slightly lower performance. This is because DF-ERC primarily focuses on multimodal inputs and lacks complex structures specifically tailored for unimodal inputs, which slightly limits its performance. Nonetheless, incorporating audio and video features into the model with our efficient fusion techniques (i.e., CFM and CRM) leads to a significant improvement in performance.

4.4 In-depth Analysis

To further investigate the effectiveness of DF-ERC, we conduct in-depth analyses to answer the following questions:

Q1: How does utterance-level feature disentanglement influence the feature fusion process? We analyzed the influence of utterance-level disentanglement on context weight and final performance. As shown in Figure 6, we divided all instances into five groups based on the similarities between modalities, sorted in ascending order. From the similarity curve, we found that the use of utterance-level disentanglement can significantly improve the similarities between modalities within an utterance, demonstrating that it effectively captures utterance-level information. Furthermore, we observed that utterance-level disentanglement is more effective in the case of utterances with a lower similarity between modalities (demonstrated by the larger F1 score gap). This is because low similarity often indicates that the features of the utterance are not fully exploited, and adding utterance-level disentanglement brings the utterance-level distance closer, thus improving similarity to a greater extent. Finally, considering the F1 metric, utterance-level disentanglement contributes more to the performance of utterances with low similarities, indicating that it can improve the final performance by leveraging utterance-level similarity.

Q2: How does modality-level feature disentanglement influence the feature fusion process? We analyzed the effect of modality-level feature disentanglement on the weight of each modality and the final F1 score. From the curves depicted in Figure 7, it is evident that integrating modality disentanglement resulted in increased weights for the video and audio modalities, especially in

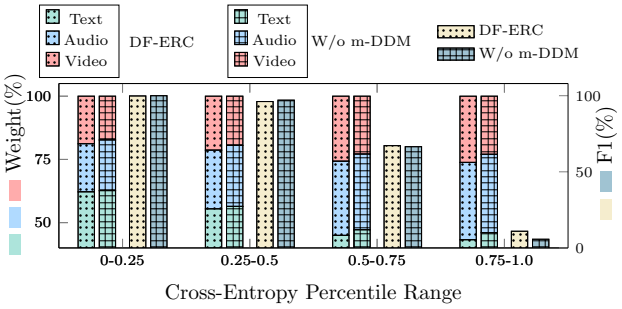
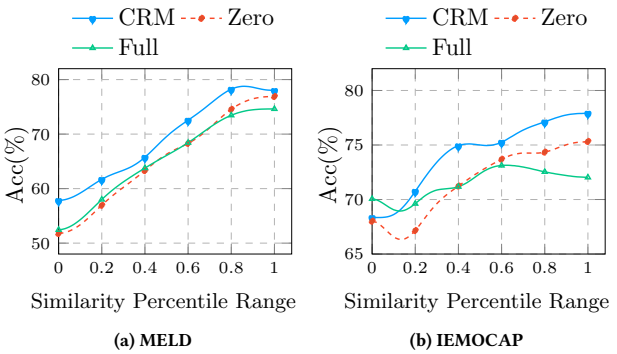


Figure 7: The influence of modality-level disentanglement (m-DDM) on the weight of different modalities and the overall performance. The x-axis represents the utterances sorted by their cross-entropy values in ascending order and divided into four groups based on percentiles (25th, 50th, and 75th percentiles). The dual y-axis shows the average weight of each modality within each group (left) and the corresponding F1 score (right). The equation for calculating the weight of each modality within an utterance is given by Eq. (12).



(a) MELD **(b) IEMOCAP**

Figure 8: Correlation between performance and modality consistency, where a larger value of the x-axis represents a higher modality consistency. We use similarity to represent consistency, where a higher similarity of different modalities indicates a higher consistency. CRM: Our proposed Context Refusion Mechanism dynamically determines how much contextual features are used based on modality consistency; Zero: using no contextual features; Full: using 100% contextual features.

utterances with higher cross-entropy values. This can be attributed to the fact that modality disentanglement enables each modality's unique characteristics to be fully exploited, resulting in a subtle enhancement of weaker modalities, such as video and audio. Moreover, this increase of weight for weaker modalities is more evident in utterances with suboptimal prediction results, where the text modality does not perform well. We observed the most substantial improvement in the F1 score for utterances with poorer prediction results, approximately 5 points improvement for those with cross-entropy ranking percentile > 0.75. This suggests that the integration of modality-level feature disentanglement can lead to more accurate predictions in challenging situations.

Q3: Do the CRM context weights decided by modality consistency really work? To verify the effectiveness of the modality similarity comparison module, we study the relationship between prediction performance and modality similarity. Additionally, we include the performance under full weight (incorporating all context representations) and zero weight (excluding all context representations). Figure 8 displays these results, with the x-axis representing the instance's similarity score ranking among all instances, while larger x values denote higher similarity. Firstly, we find that as modality similarity improves, the overall performance of the model also increases gradually. This is because the more similar the three modalities are, the more consistency they exhibit in emotion recognition, resulting in higher prediction scores. Secondly, when comparing full weight to zero weight, we observe that for utterances with relatively small similarities, the performance of full weight is superior to that of zero weight. This is because when the discrepancy between different modalities within an utterance is substantial, introducing context utterance representations can help to better recognize emotions. Conversely, as similarity increases and the discrepancy between different modalities diminishes, context representations interfere with emotional judgments. Thus, for the latter half of the figures, the performance score of zero weight outperforms that of full weight. Finally, regardless of whether full weight or zero weight is used, both approaches have limitations in that they can not be flexibly adjusted as the consistency of modalities. Our CRM can adjust context weight according to the consistency between modalities, achieving the best performance in all instances and verifying that DF-ERC effectively captures the relationship between multimodal features and context features.

5 CONCLUSION

In this work, we introduce a novel MM-ERC system that emphasizes both feature disentanglement and fusion while taking into account both multimodalities and conversational contexts. Our proposed Dual-level Disentanglement Mechanism (DDM) successfully disentangles modality- and utterance-level features using contrastive learning, while the Contribution-aware Fusion Mechanism (CFM) and Context Refusion Mechanism (CRM) fuse multimodal and contextual features effectively. Extensive experiments on two public datasets demonstrate that DF-ERC achieves the best performance compared with 19 models. Ablation studies and in-depth analyses substantiate the rationality of our approaches for controllable fusing multimodal and context features. Intuitively, our proposed approaches are not limited to emotion recognition in dialogs, and we will evaluate them on other multimodal tasks in the future.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (Grant No. 2022YFB3103602, Grant No. 2017YFC1200500), the National Natural Science Foundation of China (Grant No. 62176187), China Scholarship Council (CSC), NExT Research Center, and the Research Foundation of Ministry of Education of China (Grant No. 18JZD015).

REFERENCES

[1] Md Shad Akhtar, Dushyant Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis. In *Proceedings of NAACL-HLT*. Association for Computational Linguistics, Minneapolis, Minnesota, 370–379. https://doi.org/10.18653/v1/N19-1034

[2] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of ACL*. Association for Computational Linguistics, Melbourne, Australia, 2236–2246. https://doi.org/10.18653/v1/P18-1208

[3] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE TPAMI* 41, 2 (feb 2019), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

[4] Keshav Bansal, Harsh Agarwal, Abhinav Joshi, and Ashutosh Modi. 2022. Shapes of Emotions: Multimodal Emotion Recognition in Conversations via Emotion Shifts. In *Workshop on COLING*. International Conference on Computational Linguistics, Virtual, 44–56. https://aclanthology.org/2022.mmmpie-1.6

[5] Emad Barsoum, Cha Zhang, Cristian Canton-Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of ICMI*. ACM, New York, NY, USA, 279–283. https://doi.org/10.1145/2993148.2993165

[6] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation* 42, 4 (2008), 335–359. https://doi.org/10.1007/s10579-008-9076-6

[7] Dushyant Singh Chauhan, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Context-aware Interactive Attention for Multi-modal Sentiment and Emotion Analysis. In *Proceedings of EMNLP*. Association for Computational Linguistics, Hong Kong, China, 5647–5657. https://doi.org/10.18653/v1/D19-1566

[8] Feiyang Chen, Ziqian Luo, Yanyan Xu, and Dengfeng Ke. 2020. Complementary Fusion of Multi-Features and Multi-Modalities in Sentiment Analysis. In *Proceedings of Workshop Affective Content Analysis with AAAI (CEUR Workshop Proceedings, Vol. 2614)*. CEUR-WS.org, New York, USA, 82–99. https://ceur-ws.org/Vol-2614/AffCon20_session1_complementary.pdf

[9] Feiyu Chen, Zhengxiao Sun, Deqiang Ouyang, Xueliang Liu, and Jie Shao. 2021. Learning What and When to Drop: Adaptive Multimodal and Contextual Dynamics for Emotion Recognition in Conversation. In *Proceedings of ACM MM*. ACM, Virtual Event, China, 1064–1073. https://doi.org/10.1145/3474085.3475661

[10] Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. 2022. M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation. In *IEEE/CVF CVPR Workshops*. IEEE, New Orleans, LA, USA, 4651–4660. https://doi.org/10.1109/CVPRW56347.2022.00511

[11] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. 2019. Addressing Failure Prediction by Learning Model Confidence. In *Proceedings of NeurIPS*. IEEE, Vancouver, BC, Canada, 2898–2909. https://proceedings.neurips.cc/paper/2019/hash/757f843a169cc678064d9530d12a1881-Abstract.html

[12] Ringki Das and Thoudam Doren Singh. 2023. Multimodal Sentiment Analysis: A Survey of Methods, Trends and Challenges. *ACM Comput. Surv.* 55, 13s (mar 2023), 38 pages. https://doi.org/10.1145/3586075

[13] Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving Empathetic Response Generation by Recognizing Emotion Cause in Conversations. In *Findings of EMNLP*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 807–819. https://doi.org/10.18653/v1/2021.findings-emnlp.70

[14] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of EMNLP*. Association for Computational Linguistics, Virtual Event / Punta Cana, Dominican Republic, 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552

[15] Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual Inter-modal Attention for Multi-modal Sentiment Analysis. In *Proceedings of EMNLP*. Association for Computational Linguistics, Brussels, Belgium, 3454–3466. https://doi.org/10.18653/v1/D18-1382

[16] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In *Findings of EMNLP*. Association for Computational Linguistics, Online, 2470–2481. https://doi.org/10.18653/v1/2020.findings-emnlp.224

[17] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *Proceedings of EMNLP-IJCNLP*. Association for Computational Linguistics, Hong Kong, China, 154–164. https://doi.org/10.18653/v1/D19-1015

[18] Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In *Proceedings of EMNLP*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 9180–9192. https://doi.org/10.18653/v1/2021.emnlp-main.723

[19] Zongbo Han, Fan Yang, Junzhou Huang, Changqing Zhang, and Jianhua Yao. 2022. Multimodal Dynamics: Dynamical Fusion for Trustworthy Multimodal Classification. In *IEEE/CVF CVPR*. IEEE, New Orleans, LA, USA, 20675–20685. https://doi.org/10.1109/CVPR52688.2022.02005

[20] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *Proceedings of ACM MM*. ACM, Virtual Event / Seattle, WA, USA, 1122–1131. https://doi.org/10.1145/3394171.3413678

[21] Dou Hu, Xiaolong Hou, Lingwei Wei, Lian-Xin Jiang, and Yang Mo. 2022. MM-DFN: Multimodal Dynamic Fusion Network for Emotion Recognition in Conversations. In *IEEE ICASSP*. IEEE, Virtual and Singapore, 7037–7041. https://doi.org/10.1109/ICASSP43922.2022.9747397

[22] Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. In *Proceedings of ACL*. Association for Computational Linguistics, Online, 7042–7052. https://doi.org/10.18653/v1/2021.acl-long.547

[23] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. In *Proceedings of EMNLP*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 7837–7851. https://aclanthology.org/2022.emnlp-main.534

[24] Jiaxiong Hu, Yun Huang, Xiaozhu Hu, and Yingqing Xu. 2023. The Acoustically Emotion-Aware Conversational Agent With Speech Emotion Recognition and Empathetic Responses. *IEEE Trans. Affect. Comput.* 14, 1 (2023), 17–30. https://doi.org/10.1109/TAFFC.2022.3205919

[25] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. In *Proceedings of ACL*. Association for Computational Linguistics, Online, 5666–5675. https://doi.org/10.18653/v1/2021.acl-long.440

[26] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *IEEE Conference on CVPR*. IEEE Computer Society, Honolulu, HI, USA, 2261–2269. https://doi.org/10.1109/CVPR.2017.243

[27] Jian Huang, Zehang Lin, Zhenguo Yang, and Wenyin Liu. 2021. Temporal Graph Convolutional Network for Multimodal Sentiment Analysis. In *Proceedings of ICMI*. ACM, Montréal, QC, Canada, 239–247. https://doi.org/10.1145/3462244.3479939

[28] Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. Relation-aware Graph Attention Networks with Relational Position Encodings for Emotion Recognition in Conversations. In *Proceedings of EMNLP*. Association for Computational Linguistics, Online, 7360–7370. https://doi.org/10.18653/v1/2020.emnlp-main.597

[29] Summaira Jabeen, Xi Li, Muhammad Shoib Amin, Omar Bourahla, Songyuan Li, and Abdul Jabbar. 2023. A Review on Methods and Applications in Multimodal Deep Learning. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 2s, Article 76 (feb 2023), 41 pages. https://doi.org/10.1145/3545572

[30] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Trans. Assoc. Comput. Linguistics* 8 (2020), 64–77. https://doi.org/10.1162/tacl_a_00300

[31] Bongseok Lee and Yong Suk Choi. 2021. Graph Based Network with Contextualized Representations of Turns in Dialogue. In *Proceedings of EMNLP*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 443–455. https://doi.org/10.18653/v1/2021.emnlp-main.36

[32] Chi-Chun Lee, Carlos Busso, Sungbok Lee, and Shrikanth S. Narayanan. 2009. Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions. In *INTERSPEECH*. ISCA, Brighton, United Kingdom, 1983–1986. http://www.isca-speech.org/archive/interspeech_2009/i09_1983.html

[33] Joosung Lee and Wooin Lee. 2022. CoMPM: Context Modeling with Speaker's Pre-trained Memory Tracking for Emotion Recognition in Conversation. In *Proceedings of NAACL-HLT*. Association for Computational Linguistics, Seattle, United States, 5669–5679. https://doi.org/10.18653/v1/2022.naacl-main.416

[34] Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020. HiTrans: A Transformer-Based Context- and Speaker-Sensitive Model for Emotion Detection in Conversations. In *Proceedings of COLING*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 4190–4200. https://doi.org/10.18653/v1/2020.coling-main.370

[35] Jiangnan Li, Zheng Lin, Peng Fu, Qingyi Si, and Weiping Wang. 2020. A Hierarchical Transformer with Speaker Modeling for Emotion Recognition in Conversation. *CoRR* abs/2012.14781 (2020). arXiv:2012.14781 https://arxiv.org/abs/2012.14781

[36] Zheng Lian, Bin Liu, and Jianhua Tao. 2021. CTNet: Conversational Transformer Network for Emotion Recognition. *IEEE ACM Trans. Audio Speech Lang. Process.* 29 (2021), 985–1000. https://doi.org/10.1109/TASLP.2021.3049898

[37] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[38] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of ICLR*. OpenReview.net, New Orleans, LA. https://openreview.net/forum?id=Bkg6RiCqY7

[39] Sijie Mai, Haifeng Hu, and Songlong Xing. 2020. Modality to Modality Translation: An Adversarial Representation Learning and Graph Fusion Network for Multimodal Fusion. In *Proceedings of AAAI*. AAAI Press, New York, USA, 164–172. https://ojs.aaai.org/index.php/AAAI/article/view/5347

[40] François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *J. Artif. Intell. Res.* 30 (2007), 457–500. https://doi.org/10.1613/jair.2349

[41] Navonil Majumder, Devamanyu Hazarika, Alexander F. Gelbukh, Erik Cambria, and Soujanya Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowl. Based Syst.* 161 (2018), 124–133. https://doi.org/10.1016/j.knosys.2018.07.041

[42] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In *Proceedings of AAAI*. AAAI Press, Honolulu, Hawaii, USA, 6818–6825. https://doi.org/10.1609/aaai.v33i01.33016818

[43] Yuzhao Mao, Guang Liu, Xiaojie Wang, Weiguo Gao, and Xuan Li. 2021. DialogueTRM: Exploring Multi-Modal Emotional Dynamics in a Conversation. In *Findings of EMNLP*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 2694–2704. https://doi.org/10.18653/v1/2021.findings-emnlp.229

[44] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web. In *Proceedings of ICMI* (Alicante, Spain) *(ICMI '11)*. Association for Computing Machinery, New York, NY, USA, 169–176. https://doi.org/10.1145/2070481.2070509

[45] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced Multimodal Learning via On-the-fly Gradient Modulation. In *Proceedings of CVPR*. IEEE, 8228–8237. https://doi.org/10.1109/CVPR52688.2022.00806

[46] Patrícia Pereira, Helena Moniz, and João Paulo Carvalho. 2022. Deep Emotion Recognition in Textual Conversations: A Survey. *CoRR* abs/2211.09172 (2022). https://doi.org/10.48550/arXiv.2211.09172 arXiv:2211.09172

[47] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of ACL*. Association for Computational Linguistics, Florence, Italy, 527–536. https://doi.org/10.18653/v1/p19-1050

[48] Tulika Saha, Vaibhav Gakhreja, Anindya Sundar Das, Souhitya Chakraborty, and Sriparna Saha. 2022. Towards Motivational and Empathetic Response Generation in Online Mental Health Support. In *Proceedings of ACM SIGIR* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2650–2656. https://doi.org/10.1145/3477495.3531912

[49] Björn W. Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Commun.* 53, 9-10 (2011), 1062–1087. https://doi.org/10.1016/j.specom.2011.01.011

[50] Björn W. Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wöllmer, André Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll. 2010. Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies. *IEEE Trans. Affect. Comput.* 1, 2 (2010), 119–131. https://doi.org/10.1109/T-AFFC.2010.8

[51] Xiaohui Song, Liangjun Zang, Rong Zhang, Songlin Hu, and Longtao Huang. 2022. Emotionflow: Capture the Dialogue Level Emotion Transitions. In *IEEE, ICASSP*. IEEE, Virtual and Singapore, 8542–8546. https://doi.org/10.1109/ICASSP43922.2022.9746464

[52] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of ACL*. Association for Computational Linguistics, Florence, Italy, 6558–6569. https://doi.org/10.18653/v1/P19-1656

[53] Geng Tu, Bin Liang, Dazhi Jiang, and Ruifeng Xu. 2022. Sentiment- Emotion- and Context-guided Knowledge Selection Framework for Emotion Recognition in Conversations. *IEEE Transactions on Affective Computing* (2022), 1–14. https://doi.org/10.1109/TAFFC.2022.3223517

[54] Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605.

[55] Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A Large-Scale Dataset for Empathetic Response Generation. In *Proceedings of EMNLP*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 1251–1264. https://doi.org/10.18653/v1/2021.emnlp-main.96

[56] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context. *IEEE Intelligent Systems* 28 (2013), 46–53.

[57] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled Representation Learning for Multimodal Emotion Recognition. In *Proceedings of ACM MM* (Lisboa, Portugal). Association for Computing Machinery, New York, NY, USA, 1642–1651. https://doi.org/10.1145/3503161.3547754

[58] Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. 2021. Multimodal Sentiment Detection Based on Multi-channel Graph Neural Networks. In *Proceedings of ACL*. Association for Computational Linguistics, Online, 328–339. https://doi.org/10.18653/v1/2021.acl-long.28

[59] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis. In *Proceedings of AAAI*. AAAI Press, Virtual Event, 10790–10797. https://ojs.aaai.org/index.php/AAAI/article/view/17289

[60] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of EMNLP*. Association for Computational Linguistics, Copenhagen, Denmark, 1103–1114. https://doi.org/10.18653/v1/d17-1115

[61] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory Fusion Network for Multi-view Sequential Learning. In *Proceedings of AAAI*. AAAI Press, New Orleans, Louisiana, USA, 5634–5641. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17341

[62] Huaizheng Zhang, Linsen Dong, Guanyu Gao, Han Hu, Yonggang Wen, and Kyle Guan. 2020. DeepQoE: A Multimodal Learning Framework for Video Quality of Experience (QoE) Prediction. *IEEE Trans. Multim.* 22, 12 (2020), 3210–3223. https://doi.org/10.1109/TMM.2020.2973828

[63] Weixiang Zhao, Yanyan Zhao, and Bing Qin. 2022. MuCDN: Mutual Conversational Detachment Network for Emotion Recognition in Multi-Party Conversations. In *Proceedings of COLING*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 7020–7030. https://aclanthology.org/2022.coling-1.612

[64] Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection. In *Proceedings of ACL*. Association for Computational Linguistics, Online, 1571–1582. https://doi.org/10.18653/v1/2021.acl-long.125

A DATASET & EXPERIMENT DETAILS

A.1 Dataset

We provide detailed descriptions of the two datasets used in this study: MELD and IEMOCAP.

MELD. MELD is a multi-party dialogue dataset consisting of conversation snippets from the TV show *Friends*. The dataset includes 1,433 dialogues and 13,708 utterances, with an average of 9.6 turns per dialogue, and features 378 unique speakers. Each utterance in the dataset is labeled with one of seven emotions, namely *joy*, *sadness*, *neutral*, *surprise*, *anger*, *fear*, and *disgust*, based on the emotion conveyed by the speaker. The detailed statistics for the MELD dataset are provided in Table 4.

Table 4: The statistics for the MELD dataset. In the dataset, there are seven different types of emotions: Neutral, Surprise, Fear, Sadness, Joy, Disgust, and Anger.

	Neut	Surp	Fea	Sad	Joy	Dis	Ang	Total
Train	4,710	1,205	268	271	683	1,743	1,109	9,989
Valid	470	150	40	22	111	163	153	1,109
Test	1,256	281	50	68	208	402	345	2,610
Total	6,436	1,636	358	361	1,002	2,308	1,607	13,708

IEMOCAP. IEMOCAP is another dataset used in this study and comprises a total of 151 dialogues and 7,433 utterances. The dataset features two speakers interacting in each session, with a total of 10 speakers across all dialogues. Each utterance in the dataset is labeled with one of six emotions: *happy*, *sad*, *neutral*, *angry*, *excited*, or *frustrated*. The detailed statistics for the IEMOCAP dataset are provided in Table 5.

Table 5: The statistics for the IEMOCAP dataset. In the dataset, there are six different types of emotions: Happy, Sad, Neutral, Angry, Excited, and Frustrated.

	Hap	Sad	Neut	Ang	Exci	Frus	Total
Train	448	736	1,229	834	653	1,346	5,246
Valid	56	103	95	99	89	122	564
Test	144	245	384	170	299	381	1,623
Total	648	1,084	1,708	1,103	1,041	1,849	7,433

A.2 Settings

We employ RoBERTa-Large to encode text content without any additional pre-processing operations. We utilize the AdamW [38] optimizer and LR scheduler with a warm-up mechanism for parameter optimization. The learning rates for the PLM layer and non-PLM layer are set to 1e-5 and 1e-3, respectively. Utterances exceeding 256 tokens are clipped to meet the model's input length requirement and reduce memory usage. Furthermore, we apply a dropout layer with a rate of 0.2 after the encoder to further enhance the performance of our model. We set the batch size to 8 and 4 for the MELD and IEMOCAP datasets, respectively. We set the temperature in DDM to 0.5 for Eq. (3) and 0.3 for Eq. (5). All experiments are conducted on Ubuntu systems with two RTX A5000 GPUs. Additional parameters are shown in Table 6.

B ADDITIONAL EXPERIMENT

In this section, we present more experiment results to investigate the performance of DF-ERC.

Table 6: Hyperparameters setting.

Parameter/Module	MELD	IEMOCAP
Encoder		
Text Embedding Dim.	1024	1024
Audio Embedding Dim.	300	1582
Video Embedding Dim.	342	342
DDM		
$\text{MLP}^{m}_{t/a/v}$ Output Dim. (Eq. 4)	300	300
$\text{MLP}^{u}_{t/a/v}$ Output Dim. (Eq. 6)	300	300
CFM		
$\text{MLP}^{r}_{t/a/v}$ Output Dim. (Eq. 19)	500	500
CRM		
BiLSTM Hidden Dim. (Eq. 23)	300	300
β (Eq. 21)	0.1	0.1
Training		
Epoch size	10	10
Max grad norm	1.0	1.0
Warmup steps	100	100
Weight Decay	0.01	0.01

Hyper-parameter Analysis In our study, we introduce certain hyperparameters and adopt a grid search strategy for their optimization. An example of these parameters is α_3, as referenced in Eq.(25). To assess the effect of the parameter and evaluate the robustness of DF-ERC, we document the variation in the F1 score as α_3 is adjusted within a particular range. As depicted in Figure 9, the F1 score fluctuates slightly with changes in α_3, peaking when $\alpha_3 = 0.3$. Performance experiences a minor decline when α_3 deviates from this optimal value, illustrating the robustness of DF-ERC around $\alpha_3 = 0.3$.

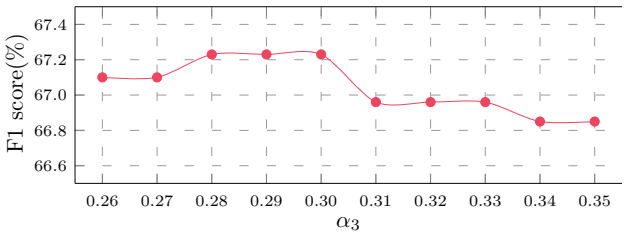
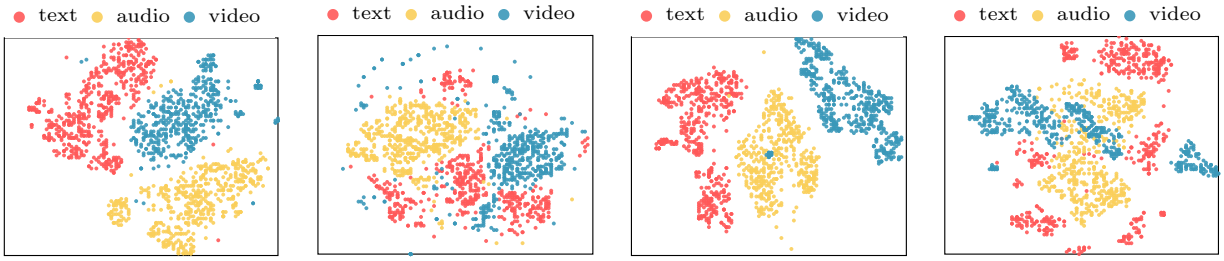


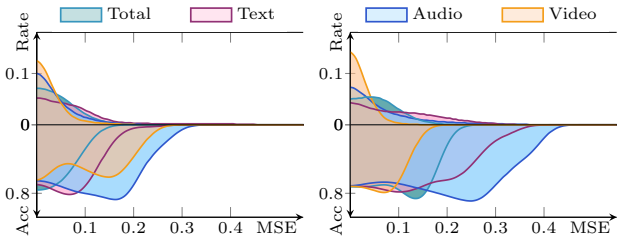
Figure 9: Trend in F1 score with respect to the changes in α_3 in Eq. (25).

Visualization for feature disentanglement. To visualize the effectiveness of DDM for feature disentanglement, we analyze the distribution of the three modalities after modality-level disentanglement (see Eq. (2)) and utterance-level disentanglement (see Eq. (4)) using t-SNE [54], as shown in Figure 10. The result



(a) Modality-MELD (b) Utterance-MELD (c) Modality-IEMOCAP (d) Utterance-IEMOCAP

Figure 10: T-SNE [54] visualization of multimodal features after applying DDM in modality and utterance levels.



(a) MELD (b) IEMOCAP

Figure 11: Virtualization of the correlation between prediction performance and MSE (Eq. (14), the difference between TCP and modality contribution) in CFM. X-axis: MSE; Upper Y-axis: Proportions of MSE; Lower Y-axis: Predicted accuracy using each modality.

indicates that modality-level contrastive learning effectively disentangles the three modalities from each other. Furthermore, we also observe that utterance-level disentanglement can align features within an utterance by entangling features from three distinct modality spaces. These findings highlight the effectiveness of our DDM in controlling the modality distribution in feature space based on the corresponding optimization objective and thus can further improve the emotion recognition performance.

The contribution of CFM for final performance. To verify the effect of CFM, we investigated the relationship between the prediction effect (evaluated using MSE) of TCP and the final emotion recognition performance. Figure 11 shows that for the majority of utterances, the MSE of the TCP prediction is less than 0.1, indicating satisfactory performance for TCP. However, it should be noted that a better TCP prediction for each modality does not necessarily result in a higher emotion prediction score, as shown by the wave pattern in the bottom half of the figure, which suggests that TCP has some limitations as a modality contribution evaluator. Nonetheless, after averaging the prediction MSE of the three modalities within an utterance, we observed a gradual increase in performance with a smoother tendency as the MSE decreased. This demonstrates that, overall, a predicted contribution weight can guide the model to assign the optimal weight for each modality and thus achieve better final performance.